Introduction to Bioinformatics
9th practice
Gene and protein annotation

Today you will practice how to use Hidden Markov Models to find and annotate genes.

The required software is only a console as described in the document pract09_connect_to_cloud_windows.pdf

Files you need:

- **All necessary program packages (glimmer, hmmer, blast) and files are available on the server**

Exercises

You have found a new unknown protein. Your task is to figure out what this protein could do and what the function of that particular protein could be.
One possible way is querying the sequence against a database with a specific program such as blast. The HMMs offer another efficient solution. In that model you compare the sequence against a functional group rather than to individual proteins.
The main steps of such an analysis are:

1. Creating a profile database (functional groups)
2. Comparing with the database
3. Evaluating the results

An example for such database is the Cluster of Orthologous Groups (COG) in which the sequences carrying out the same functions are grouped together. The functions are organized in a hierarchical way.

In the first step we are going to work with following groups:

1. `COG5274 Cytochrome b involved in lipid metabolism`
2. `COG5518 Bacteriophage capsid portal protein`
3. `COG5661 Predicted secreted Zn-dependent protease`

All the names are listed in the file:
(`/gfs/data/genome_annotation/cognames2003-2014. tab`)
Use *less* command to see the content and grep command to find record (i.e. grep `'COG5274\|COG5518\|COG5661'`
`/gfs/data/genome_annotation /cognames2003-2014.tab`)

Building hmm profile:

The first step is creating a multiple alignment from the sequences, using the clustal omega tool.
The three COG sequences are located in the '/gfs/data/genome_annotation/' directory (note that the output should be in stockholm format):

```
clustalo -i /gfs/data/genome_annotation/COG5274.fasta -o COG5274.sto --outfmt st --force
```

How does the alignment file look like?

The next step is creating the HMM profile. The program is going to estimate the proper transition and emission probabilities based on the multiple alignment:

```
hmmbuild COG5274.hmm COG5274.sto
```

Do the alignment of the hmm profile for the other two COG (COG5518.fasta, COG5661.fasta)!

Indexing the database:
The next step is building a database from the profile. In order to do that one should simply concatenate all the three hmm files:

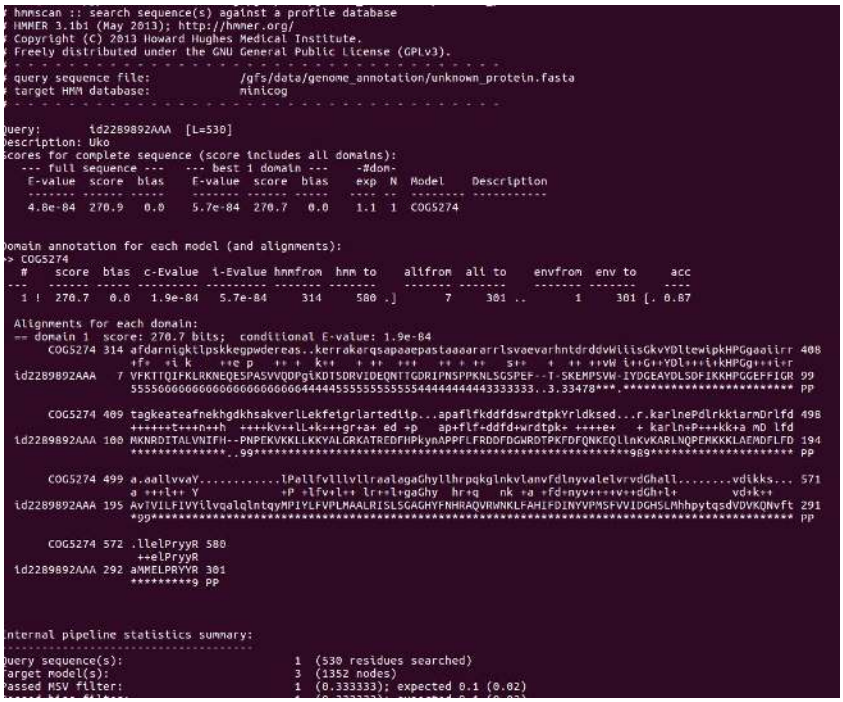
```
cat *.hmm > minicog
```

Then the database should be indexed:

```
hmmpress minicog
```

Query the database:
If the database is ready then the hmm scan could be used for comparing our unknown sequence with profile database using the hmmscan program. Path to our unknown protein is:

```
#searching database
hmmscan minicog /gfs/data/genome_annotation/unknown_protein.fasta
```



The result shows a strongly significant match. What is your suggestion, what the function of that particular protein could be?

Annotation:

Annotate the following protein using the precompiled database PFAM (Pfam-A - manually annotated)! Path to the pfam database (already built and indexed):
`/gfs/data/genome_annotation/Pfam-A.hmm`

`Path to the query sequence: /gfs/data/genome_annotation/seq1.fasta`
>unannotated
MTDTQQTVYVVEDDEAVRDSLELLLKSDGKPVKTYDNANAFLKDYSEKMAGCIVLDIRMP
GMDGMELQKKLNEKHSILPIIFVTGHGDVPMAVDAMKEGAVDFIQKPYREEALLQKIEAA
LEQDKEQRKTLGEKQEILRRVKSLTPREHEIMDRMIAGQANKVIAIELEISQRTVEIHRS
RVMHKMGTHSLAHLVRMVLSVKDLIDAR

Genome annotation

The genome annotation is usually one of the steps in an NGS (next generation sequencing) analysis pipeline.
In our case let us work with
As was mentioned in the lecture the GLIMMER program (program for finding genes in genomes) implements an interpolated Markov Model for finding genes.
The first step is finding the ORF regions (usually the long ones):
`long-orfs -n -t 1.15`
`/gfs/data/genome_annotation/helico_contigs.fasta helico.longorfs`

Selecting these regions and putting them into a fasta file:

```
extract -t  /gfs/data/genome_annotation/helico_contigs.fasta
helico.longorfs > helico.train
```

These sequences will be the training set. In that case the training set is small indicating the poor quality of the assembly.

Calculating the “context”:

```
build-icm helico.icm < helico.train
```

In these phase the program estimates the conditional probabilities based on the training set.

Finding the genes with the following settings:

1. Allowing 50 long overlap between genes
2. The minimum gene length is 110
3. Minimum threshold score (only genes having a score at least 30 are reported)

```
glimmer3 -o50 -g110 -t30
/gfs/data/genome_annotation/helico_contigs.fasta  helico.icm
helico
```

Select a predicted gene from the output of the glimmer training set (`helico.train`) and try to find what kind of domains it contains using the PFAM database (use the hmmscan).
You can translate the nucleotid sequence into proteins using online tools (i.e. https://www.ebi.ac.uk/Tools/st/emboss_transeq/ .

You may also use the BLAST programs (mostly blastx) to annotate the ORF sequences.
The indexed swissprot/uniprot database is deposited here:

```
/gfs/data/genome_annotation/uniprot_sprot.pep
```

Running the blastx:

```
blastx -query helico.train -db
/gfs/data/genome_annotation/uniprot_sprot.pep -num_threads 4
-outfmt 6 > blastx.outfmt6
```

What do you think what could be the functions of the unknown ORFs?